Analysis of Economics Data
Chapter 10: Data Summary with Multiple Regression

© A. Colin Cameron
Univ. of Calif. Davis

October 2022

CHAPTER 10: Data Summary with Multiple Regression

- Consider the relationship between house price and several variables
  - size, number of bedrooms, ...
- Mostly a straight-forward extension of bivariate regression.
- New is:
  - rely less on visual methods
  - no easy formulas for estimates (without matrix algebra)
  - adjusted R^2
  - simultaneous tests of several hypotheses (in next chapter).

Outline

1. Example: House price and characteristics
2. Two-way Scatter Plots
3. Correlation
4. Regression line
5. Interpretation of Slope Coefficients
6. Model Fit
7. Computer Output Following Multiple Regression
8. Inestimable Models

10.1 Example: House Price

- HOUSE data: 29 houses sold in central Davis, California, in 1999.
  - lot size is 1 for small, 2 for medium and 3 for large
  - a half bathroom is a lavatory without bath or shower.

Variable	Definition	Mean	Standard deviation	Min	Max
Price	Sale Price in dollars	253910	37391	204000	375000
Size	House size in square feet	1883	398	1400	3300
Bedrooms	Number of bedrooms	3.79	0.68	3	6
Bathrooms	Number of bathrooms	2.21	0.34	2	3
Lotsize	Size of lot (1, 2 or 3)	2.14	0.69	1	3
Age	House age in years	36.4	7.12	23	51
Month Sold	Month of year house was sold	5.97	1.68	3	8

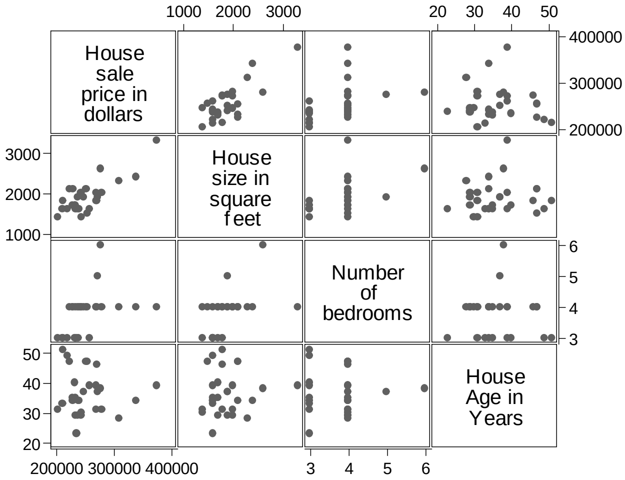
Example Regression

Variable	Coefficient	St. Error	t statistic	p value	95% conf. int.	
Size	68.37	15.39	4.44	0.000	36.45	101.29
Bedrooms	2685	9193	0.29	0.773	-16379	21749
Bathrooms	6833	15721	0.43	0.668	-25771	39437
Lot Size	2303	7227	0.32	0.753	-12684	17290
Age	-833	719	-1.16	0.259	-2325	659
Month Sold	-2089	3521	-0.59	0.559	-9390	5213
Intercept	137791	61464	2.24	0.036	10321	265261
n	29					
F(6,22)	6.83					
p-value for F	0.0003					
R^2	0.651					
Adjusted R^2	0.555					
St. error	24936					

10.2 Two-way Scatterplots

- Can get multiple two-way scatterplots - next slide.
- Some programs provide three-way surface plots
  - e.g. price against size and number of bedrooms
  - these can be difficult to read.

Two-way Scatterplots



10.3 Correlation

- **Pairwise correlations** are very useful for exploratory analysis
  - Price is most highly correlated with square feet, then bedrooms and bathrooms.
  - Asterisk means statistically significant correlation at significance level 0.05.

Correlation	Price	Size	Bed	Bath	Lot	Age	Mth Sold
Sale Price	1						
Size	.79*	1					
Bedrooms	.43*	.52*	1				
Bathrooms	.33	.32	.04	1			
Lot Size	.15	.11	.29	.10	1		
Age	−.07	.08	−.03	.03	−.02	1	
Month Sold	−.21	−.21	.18	−.39*	−.06	−.37	1

- Bedrooms correlated with Price but this could merely be picking up the effect of Size (Bedrooms is correlated with Size).
- Multiple regression measures role of each variable in predicting price, after controlling for the other variables.

10.4 Regression Line

- **Regression line** from regression of y on several variables $x_2, ..., x_k$ is

$$\widehat{y} = b_1 + b_2 x_2 + b_3 x_3 + \cdots + b_k x_k,$$

where

  - $\widehat{y}$ = predicted (or fitted) dependent variable
  - $x_2, .., x_k$ are regressor variables
  - $b_1, b_2, ..., b_k$ are estimated intercept and estimated slope parameters.

Least Squares Estimation

- The **residual** is

$$
\begin{aligned}
e_i &= y_i - \widehat{y}_i \\
&= y_i - b_1 - b_2 x_{2i} - b_3 x_{3i} + \cdots - b_k x_{ki}.
\end{aligned}
$$

- Estimate $b_1, b_2, \ldots, b_k$ by **least squares** (OLS: ordinary least squares) that minimizes sum of squared residuals

$$
\begin{aligned}
\textstyle\sum_{i=1}^{n} e_i^2 &= \textstyle\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \\
&= \textstyle\sum_{i=1}^{n} (y_i - b_1 - b_2 x_{2i} - b_3 x_{3i} + \cdots - b_k x_{ki})^2.
\end{aligned}
$$

- Estimates $b_1, \ldots, b_k$ solve the k **normal equations**
  - $\sum_{i=1}^{n} x_{ji}(y_i - b_1 - b_2 x_{2i} - b_3 x_{3i} - \cdots - b_k x_{ki}) = 0, \quad j = 1, \ldots, k,$
  - or $\sum_{i=1}^{n} x_{ji} e_i = 0, \quad j = 1, \ldots, k$
  - each regressor is orthogonal to the regressor
  - and the residuals sum to zero if an intercept is included.

Least Squares Estimates

- Consider the coefficient b_j of the j^{th} regressor x_j.
- The OLS coefficient b_j can be calculated by
  - bivariate regression of y on $\widetilde{x}_j$
  - where $\widetilde{x}_j = x_j - \widehat{x}_j$ is the residual from regressing x_j on an intercept and all regressors other than x_j.
- Algebraically

$$b_j = \frac{\sum_{i=1}^n \widetilde{x}_{ji}(y_i - \bar{y})}{\sum_{i=1}^n \widetilde{x}_{ji}{}^2}.$$

- So OLS coefficient measures the relationship between y and x_j after the explanatory power of x_j has been reduced by controlling for how the other regressors in the equation jointly predict x_j.
- More generally matrix algebra is used - see Appendix C.4.

10.5 Interpretation of Slope Coefficients

- b_2 measures the **partial effect** of changing x_2 **while holding all other regressors at their current values**
- Reason: increase x_2 by Δx_2. Then

$$\begin{aligned} \widehat{y}_{new} &= b_1 + b_2(x_2 + \Delta x_2) + b_3 x_3 + \cdots + b_k x_k \\ &= b_2 \Delta x_2 + b_1 + b_2 x_2 + b_3 x_3 + \cdots + b_k x_k \\ &= b_2 \Delta x_2 + \widehat{y}_{old} \end{aligned}$$

- So $\Delta \widehat{y} = b_2 \Delta x_2$ and hence **partial effect**

$$\left.\frac{\Delta \widehat{y}}{\Delta x_2}\right|_{x_3, \ldots, x_k} = b_2.$$

Estimated Total Effect

- The **total effect** on y_2 lets other features of the house change as we change x_2.
- Suppose $\widehat{y} = b_1 + b_2 x_2 + b_3 x_3$
  - changing x_2 by Δx_2 is associated with a change in x_3 of Δx_3
  - then the total effect on y of changing x_2 by Δx_2 equals $\Delta \widehat{y} = b_2 \Delta x_2 + b_3 \Delta x_3$
  - Dividing by Δx_2, the **total effect** on y_2 of changing x_2 equals

$$\left.\frac{\Delta \widehat{y}}{\Delta x_2}\right|_{Total} = b_2 + b_3 \frac{\Delta x_3}{\Delta x_2}$$

- Aside: Mechanical result for OLS
  - When regression is by OLS, the total effect on the predicted value of y when x_2 changes by one unit from a multivariate regression simply equals the slope coefficient from bivariate regression of y on x_2 alone.

Further Details

- Partial effect versus total effect
  - Often interest lies in the **partial effect** of changing one key regressor after controlling for other variables
  - e.g. size of change in earnings as education varies after controlling for age, gender, socioeconomic background.
- Calculus
  - partial effect of regressor x_j is partial derivative $\partial y/\partial x_j$.
  - total effect of regressor x_j is total derivative dy/dx_j.
- Causation
  - OLS measures association but not necessarily causation.
  - so say that a one unit change in x_j is associated with a b_j change in $\widehat{y}$ holding all other regressors constant.

10.6 Model Fit: Standard Error of the Regression

- For multiple regression the **standard error of the regression** is

$$s_e = \sqrt{\frac{1}{n-k}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}.$$

- Now division is by $n-k$, rather than $n-2$ in the bivariate case, as k degrees of freedom are lost since computation of $\widehat{y} = b_1 + b_2 x + \cdots + b_k x_k$ is based on the k estimates $b_1, ..., b_k$.
- Another name for s_e is the **root mean squared error (MSE) of the residual**.
- It is also sometimes called the **standard error of the residual**.

R-Squared

- Again Total SS = Explained SS + Residual SS.
- **R-squared** is same underlying formula as in bivariate case

$$R^2 = \frac{\text{Explained SS}}{\text{Total SS}} = \frac{\sum_{i=1}^n (\widehat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

$$R^2 = 1 - \frac{\text{Residual SS}}{\text{Total SS}} = 1 - \frac{\sum_{i=1}^n (y_i - \widehat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

  - assuming the model includes an intercept term
  - $0 \leq R^2 \leq 1$.
- R^2 equals the **fraction of the variation** in y (about $\bar{y}$) **explained** by the regressors $x_1, ..., x_k$.
- R^2 equals the **squared correlation** between y_i and $\widehat{y}_i$
  - i.e. between fitted and actual value of y.

Adjusted R-Squared

- R^2 **necessarily increases** as add regressors, since residual sum of squares decreases.
- So also use **adjusted R-squared**, denoted $\bar{R}^2$

$$\begin{aligned}\bar{R}^2 &= 1 - \frac{\text{Residual SS}/(n-k)}{\text{Total SS}/(n-1)} \\ &= 1 - \frac{\sum_{i=1}^n (y_i - \widehat{y}_i)^2/(n-k)}{\sum_{i=1}^n (y_i - \bar{y})^2/(n-1)}.\end{aligned}$$

- Motivation is to divide residual and total sum of squares by their degrees of freedom
  - this gives penalty to larger models ($k \uparrow$)
- Compare smaller and larger model for house price
  - with just square feet as regressor: $R^2 = 0.618$ and $\bar{R}^2 = 0.603$.
  - with all regressors: $R^2 = 0.651$ and $\bar{R}^2 = 0.555$.
  - only a modest increase in R^2 and $\bar{R}^2$ falls.

Information Criteria

- **Information criteria are a more advanced method that penalizes larger models.**
- Specifically, information criteria penalize $\widehat{\sigma}_e^2$ for larger model size
  - $\widehat{\sigma}_e^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \widehat{y}_i)^2$ is the sample average of the squared residuals
  - similar to s_e^2 except there is no degrees of freedom correction, so division is by n rather than $n-k$.

Criteria		**General formula**
Akaike IC	AIC	$= n \times \ln \widehat{\sigma}_e^2 + n(1+\ln 2\pi) + 2k$
Bayesian IC	BIC	$= n \times \ln \widehat{\sigma}_e^2 + n(1+\ln 2\pi) + k \times \ln(n)$
Hannan-Quinn IC	$HQIC$	$= n \times \ln \widehat{\sigma}_e^2 + n(1+\ln 2\pi) + 2k \times \ln(\ln(n))$

  - k is the number of regressors
  - smaller values of each criterion are preferred
  - BIC is preferred (AIC has too small a penalty for model size)
  - some statistical packages divide the above formulas by n.

10.7 Computer Output Following Multiple Regression

- Computer output usually has **three components**
- 1. ANOVA table
  - Gives explained, residual and total sum of squares
  - Use to compute R-squared (and overall F-statistic given in next chapter).
- 2. Regression coefficient estimates
  - and associated standard errors, t-statistics, p-values, CI's
- 3. Regression summary statistics
  - number of observations, R-squared, adjusted R-squared, Standard error of regression, overall F-statistic.

10.8 Inestimable Models

- It is not always possible to estimate all k regression coefficients in the regression of y on an intercept and regressors $x_2, \ldots, x_k$.
  - e.g. bivariate regression cannot estimate b_2 if $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0$.
- Then computer regression output will have no entries for one or more regressors, and may include the word omitted.
- When not all coefficients can be estimated
  - the **coefficients** are said to be **not identified**
  - the **regressors** are said to be **perfectly collinear**
  - the **regressor data matrix** is said to of **less than full rank.**
- This situation may arise due to
  - **inadequate variation** in the data in a well-specified model
  - or due to a **poorly specified model**.

Key Stata Commands

```
clear
use AED_HOUSE.DTA
correlate price size bedrooms bathroom lotsize age
       monthsold
regress price size bedrooms bathroom lotsize age
       monthsold
```

Some in-class Exercises

1. Regression leads to fitted line $\widehat{y} = 2 + 3x_2 + 4x_3$. What is the residual for observation $(x_2, x_3, y) = (2, 1, 9)$?
2. Suppose we know that $y = 8 + 5x_2 + 5x_3 + u$ where $E[u|x] = 0$. Give the conditional mean of y given x and the error term for the observation $(x, y) = (2, 3, 30)$.
3. OLS regression on the same dataset leads to fitted models $\widehat{y} = 6 + 5x_2$ and $\widehat{y} = 2 + 3x_2 + 4x_3$. Are you surprised by the different coefficients for x_2? Explain.
4. OLS regression of y on x for a sample of size 53 leads to residual sum of squares 20 and total sum of squares 50. Compute the standard error of the regression.
5. For the data of the previous example, compute R^2 and the correlation between y.and $\widehat{y}$.